Waqas Hassan, Philip Birch, Rupert Young, Chris Chatwin
Univ. of Sussex (United Kingdom)
Proc. SPIE 9026, Video Surveillance and Transportation Imaging Applications 2014, 902602
(March 5, 2014); doi:10.1117/12.2039181

PHACT: Parallel HOG and Correlation Tracking

Histogram of Oriented Gradients (HOG) based methods for the detection of humans have become one of the most reliable methods of detecting pedestrians with a single passive imaging camera. However, they are not 100 percent reliable. This paper presents an improved tracker for the monitoring of pedestrians within images. The Parallel HOG and Correlation Tracking (PHACT) algorithm utilises self learning to overcome the drifting problem. A detection algorithm that utilises HOG features runs in parallel to an adaptive and stateful correlator. The combination of both acting in a cascade provides a much more robust tracker than the two components separately could produce.

1. Introduction

The tracking humans in video remains a difficult problem. People move in a nonlinear and unpredictable manner, are non rigid, and there is a wide degree of variation between different people. There are many applications for such a system including security alerts, complete localisation of several people [1], counting people in groups [2] and behaviour and crowd analysis[3].

Histogram of Oriented Gradients (HOG) [4] feature descriptors have been successfully shown to be one of the most reliable methods of detecting pedestrians and other objects within an image. A histogram of the orientations of the edges is produced from the training set and this data is fed into a linear support vector machine (SVM). To locate a possible target the HOG descriptors are again calculated within a sliding window and this is tested against the SVM classifier to determine if the window contains the object. Strictly speaking the HOG method is not a full tracker. It is purely a detector.

Correlation based tracking, on the other hand, has had a great degree of success for target tracking and identification. Much of the research in recent years has focussed on composite filters that combine multiple input training images; image noise and clutter structure; and out of class images to produce a robust filter. These filters are robust to noise; intensity variations and can work in real time. Alone, the filter has no adaptivity.

The predictive component of a tracker ensures that the track remains on the correct object and is especially important when two objects cross each other. The simplest predictor is to look for the overlap between frames or find the nearest object. A more intelligent approach is to measure the objects velocity vector. Kalman filters are the classic method for achieving this. However, these are linear so they have been extended to the non-linear Extended Kalman filter and unscented Kalman [5]. Particle filters are also widely used [6]. They cope well with the changing direction of the objects but are best suited for extended objects within the image and can suffer from sampling problems. To overcome the problems encountered by the target changing direction most state of the art trackers (e.g. [7], [8] now use exhaustive search based methods, i.e., apply the descriptor to all reasonable possible locations. This paper has opted for this approach.

Several groups have attempted this task. Some of the first methods used techniques such as frame differencing, motion and colour to detect humans [9].

There have been a number of papers that look at the tracking problem alone and leave the detection to a human operator. Hassan [6] used particle filters, colour histograms, and optical flow. Yilmaz [8] and Yang [10] provide a survey of many of the major techniques. This paper is concerned with both detection and tracking.

This paper combines the HOG and correlation based tracking and presents the PHACT: parallel HOG and correlation filter, which has been designed to track humans. The basic design philosophy is that the algorithm locates all the possible objects of the class within the image and tracks them. Firstly, a HOG detector is used to locate all possible humans within the image and secondly, the correlator is no longer a fixed template, but an adaptive system. The object classification is performed using a HOG based classifier. The classifier has to be to trained off-line for the specific set of objects. The tracker is therefore not suitable for tracking any arbitrary object without a training period. However, the trained classes can be rather generic such as people or vehicles making is suitable for crime detection applications.

Once the set of objects are detected in the image frame, a rectangular set of coordinate for each object is returned. This then feeds into a correlation algorithm. The correlation peak is then detected and this is used as the final track result. This correlation mask is used in subsequent frames until the HOG detector again finds a suitable target and the mask is replaced with the new image.

Occasionally the HOG detector will produce a false positive, it could for example lock on to an area of road. This will then be fed to correlator which will then produce a very good match since it is correlating the same two images with each other. Without suppressing this, the PHACT would permanently lock onto the background. The algorithm overcomes this by comparing the HOG rectangle with a running average background image. If the HOG image correlators more strongly with the background than the current frame, the track is rejected.

The algorithm has been tested on two video sequences and performance has been evaluated against the existing HOG based method. The experimental results has shown that the tracker has a good success rate and it has shown resistance to noise, clutter and lighting and colour changes as compare to the HOG detector.

The following section explains the PHACT design. The correlation filter design is discussed in section 3 and results are shown in section 4. The study is finally concluded in section 5.

2. PHACT Design

A number of different tracking methods have been discussed in the previous section. A new category of tracker is presented in this paper where objects are tracked based on their class. We call this approach "track by class". All the objects are first located in the frame and then tracked based on the type of class they belong to. The objects can be differentiated by set of correlation filters. This approach divides our algorithm into three components:

- Object classification
- Region of support extraction
- Object detection through cross correlation

Objects are classified based on HOG base classifier. A HOG based classifier needs an offline training procedure. Once the HOG is trained it can then be used to detect specific type of objects in the scene. This means that the tracker is not suitable for tracking any arbitrary object without a training period. On the other hand this also means that the tracker is rather generic can be trained to track any class of objects like human or vehicles making it a powerful tracking tool.

Once object are detected by the HOG base classifier, a rectangular region of interest (ROI) is obtained for each object. This ROI for each object is then feeds into a correlation algorithm to train the temple for tracking that particular object. This template is then stores in the database to keep track of the object while its in the scene and not detected by the HOG detector.

Once the HOG detector finds the object again, the same process is repeated and the template in the database is replaced by the new ROI.

Occasionally the HOG detector will produce a false positive: it could for example lock on to an area of road. This will then be fed to the correlator which will then produce a very good match since it is correlating the same two images with each other. Without suppressing this, the PHACT would permanently lock onto the background. The algorithm overcomes this by comparing the HOG rectangle with a running average background image. If the HOG rectangle correlates more strongly with the background than the current frame, the track is rejected.

3. Correlation Filter Design

Several designs of correlator have been tested. The simplest is the normalised cross correlation. This can be further improved by band limiting the image by DOG filtering the templates [11]. Both of these options only work on the single previous state. The tracker can be improved further by comparing several past templates. If the nth past template is described as T_n we could test each template individually for all n:

$$C = \sum_{n=1}^{N} I * T_n$$

where $*$ is the cross-correlation operator and I is the input image. We would then look for a peak in C. This is rather computationally intensive but we note since the correlation operator is linear:

$$C = I * \sum_{n=1}^{N} T_n$$

The problem is now that there is probably a large degree of similarity between individual filters since they are from the same object, meaning that I will actually correlate against a number of the filters T_n making the value of C rather unstable for different inputs. To overcome this we can replace the multiple set of T_n with a single filter that encompasses all the individual templates and has a number of design criteria added in. This is known as a composite filter. There are a number different designs but we have incorporated the optimum trade off maximum average correlation height (OT-MACH) filter [12] due to its known performance.

The filter works by attempting maximise the average correlation height (ACH) for all the templates. It attempts to minimise the average correlation energy (ACE), the average similarity matrix (ASM) and output noise variance (ONV) of the filter.

The ASM is a measure of how similar each correlation template is to the others. By minimising it, the filter then gives the same output correlation value no matter which template the input actually matches against. Minimising the ACE forces the filter to give a sharp peak when a match is produced. The ONV is a measure of the filter's ability to reject noise and clutter. The filter in frequency space is then given by [12]

$$h = D^{-1} m'$$

where

$$D = \alpha P + \beta D_x + \gamma S_x$$

where P is the noise power spectral density, D_x is the mean power spectral density of the templates, and S_x is the absolute mean difference between the mean Fourier transform of the templates and each template, i.e. the variance of the Fourier transform of the templates. m' is the complex conjugate of the mean of the Fourier transform of the template images, T_n. D is a two dimensional array so the $^{-1}$ operator is a pixel level divide, rather than an array inversion (i.e. equivalent to a Matlab ./).

α, β, λ are tuning parameters that allow the adjustments of the discrimination of the filter and its noise rejection ability. Five past states, as produced by the HOG filter, where used to train the MACH. It is the output of this filter, i.e., the position of the peak in the correlation output, that is used as the track result. One advantage the MACH filters have

over single template filters is that since the filter is trained on multiple angles and multiple scales, a degree of out of plane rotation and scale invariance is introduced. To perform the correlation, the inverse Fourier transform is calculated of *h* and this cross-correlated in the space domain with the current frame.

4. Results

The proposed PHACT method has been tested on different video sequences and has shown promising results. When tested on the video sequence where the tracked person continuously changed position and direction, the presented PHACT method tracked the object for 99% of the time as compare to HOG detector where the detection rate was only 26%. Figure 1 shows an image from the sequence with the tracked object. The technique was also tested for multiple objects tracking with results shown in Figure 2 on a train station [13].

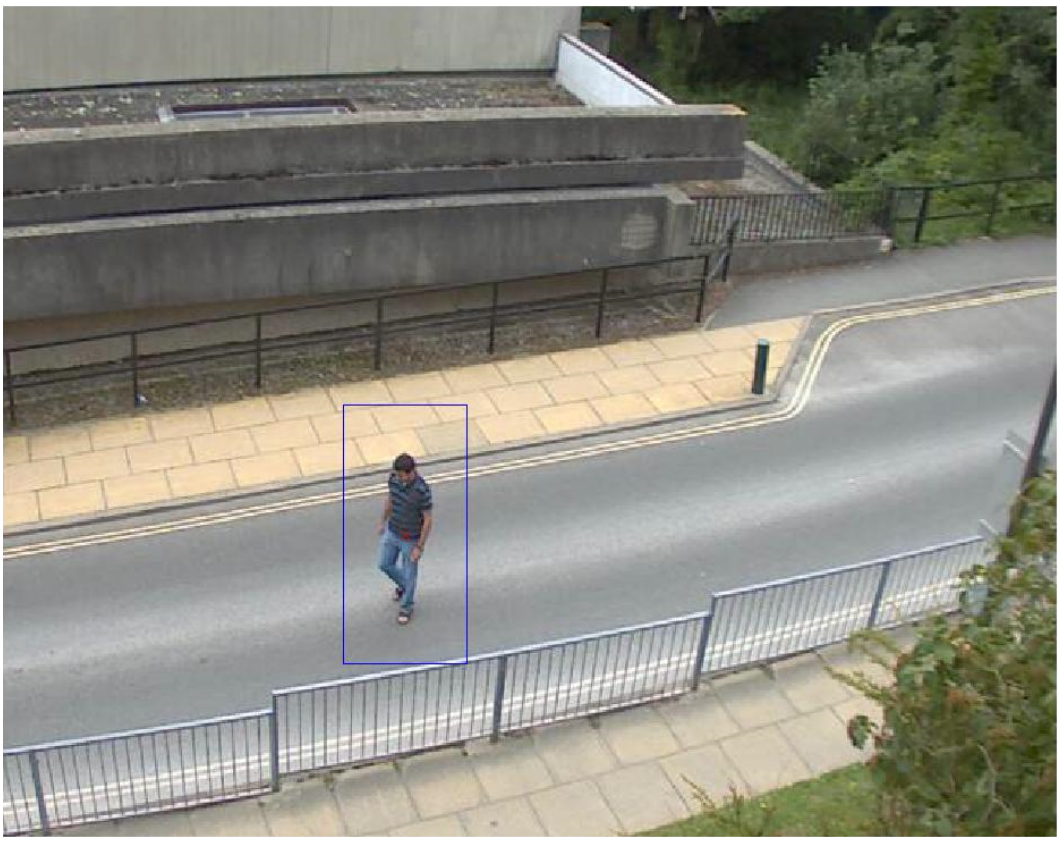

Figure 1: Scene from test video sequence. Blue boxes represent the HOG detection where the red crosses indicate the PHACT.

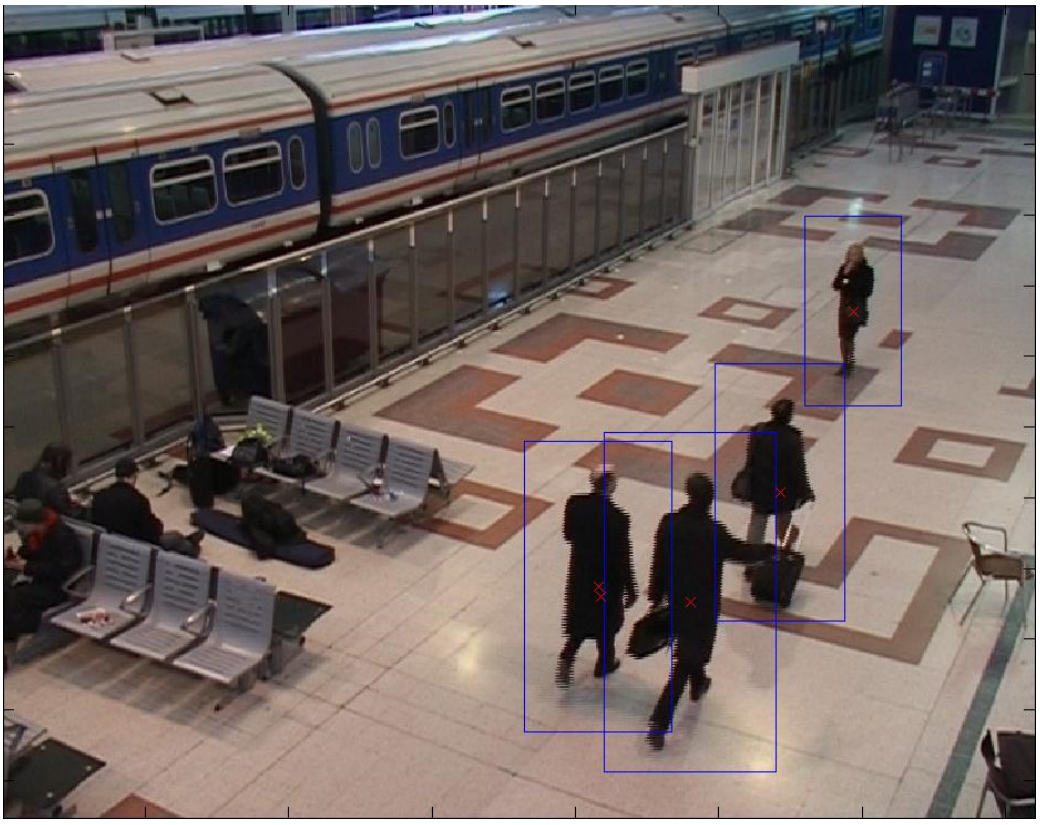

Figure 2: Scene from PETS dataset showing the working of PHACT. Blue boxes represent the HOG detection where the red crosses indicate the PHACT. Note that women walking towards the camera in the top of the scene is missed by HOG and detected by PHACT

The MACH filter has a large degree of invariance to image degradation and lighting changes. This is demonstrated below. A video sequence was recorded with a fixed exposure time and aperture whilst the lights in the room where changed (see Figure 3). The MACH filter can still determine the position of the person, whilst other techniques such as the colour histogram fail.

The experimental results clearly show that the HOG detector does not perform well on the test sequences. The proposed method that combined the HOG detector with correlation filter improves the tracking results where the results are further improved by using the combination of DOG and MACH filtering.

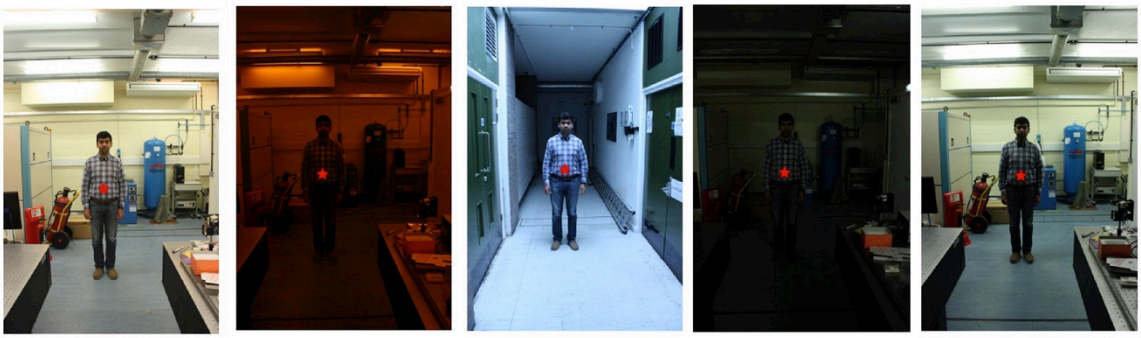

Figure 3: PHACT working in different lighting conditions.

5. Conclusion

HOG and correlation based tracking method is presented in this paper where objects are tracked based on type of class they belong to. HOG is initially trained offline to detect the object of interest. A tracking template is then generated based on the ROI of the detected object and objects are tracked. Experiment results have shown that the proposed method works at 99% of the time as compare to HOG which only works at 26% on the testing sequence. Also it has been observed that tracking results can further be improved by combining DOG and MACH filter based tracking as compare to simple correlation.

References

1. S.-I. Yu, Y. Yang, and A. Hauptmann, "Harry Potter's Marauder's Map: Localizing and Tracking Multiple Persons-of-Interest by Non- negative Discretization," *IEEE CVPR*, 2013.

2. D. Fehr, R. Sivalingam, V. Morellas, N. Papanikolopoulos, O. Lotfallah, and Y. Park, "Counting People in Groups," *Advanced Video and Signal Based Surveillance, 2009. AVSS '09. Sixth IEEE International Conference on*, pp. 152–157, 2009.

3. T. Moeslund, A. Hilton, and V. Kru¨ger, "A survey of advances in vision-based human mo- tion capture and analysis," *Computer Vision And Image Understanding*, vol. 104, no. 2-3, pp. 90–126, 2006.

4. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *IEEE Computer Society Conference on Computer Vi- sion and Pattern Recognition*, vol. 1, p. 886, 2005.

5. E. A. Wan and R. Van Der Merwe, "The unscented Kalman filter for nonlinear estima- tion," *Adaptive Systems for Signal Process- ing, Communications, and Control Symposium*, pp. 153–158, 2000.

6. W. Hassan, N. Bangalore, P. Birch, R. Young, and C. Chatwin, "An adaptive sample count particle filter," *Computer Vision And Image Understanding*, vol. 116, pp. 1208–1222, Dec. 2012.

7. J. Santner, C. Leistner, A. Saffari, T. Pock, and H. Bischof, “PROST: Parallel robust online simple tracking,” in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Con- ference on*, pp. 723–730, 2010.

8. A. Yilmaz, O. Javed, and M. Shah, “Object tracking: A survey,” *Acm Computing Surveys*, vol. 38, pp. 13–es, Dec. 2006.

9. C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, “Pfinder: real-time tracking of the human body,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, no. 7, pp. 780–785, 1997.

10. H. Yang, L. Shao, F. Zheng, L. Wang, and Z. Song, “Recent advances and trends in visual tracking: A review,” *Neurocomputing*, vol. 74, pp. 3823–3831, Nov. 2011.

11. L. Jamal-Aldin, R. Young, and C. Chatwin, “Synthetic discriminant function filter employing nonlinear space-domain preprocessing on bandpass-filtered images,” Applied Optics, vol. 37, no. 11, pp. 2051–2062, 1998.

12. A. Mahalanobis, B. Kumar, S. Song, S. Sims, and J. Epperson, “Unconstrained Correlation Filters,” Applied Optics, vol. 33, no. 17, pp. 3751–3759, 1994.

13. PETS Dataset.